

# Findings from the DIL Interviews: Data Management and Organization

## Skills in this competency may include:

- Understands the lifecycle of data, develops data management plans, and keeps track of the relation of subsets or processed data to the original data sets.
- Creates standard operating procedures for data management and documentation.

## Additional skills mentioned by interviewees:

- Familiarity with tools for data management
- Higher level of annotation on data sets to keep track of changes and analyses performed

**Average Ranking of Importance (5=essential):** Faculty=4.00, Students = 4.47

## Faculty responses:

Faculty describe data management skills as standard operating procedures that are passed down from one student to the next. Rudimentary skills are believed to be learned in statistics courses.

When it comes to the lifecycle of data, faculty have different perspectives. One believes that these skills are not necessarily tied to the life-cycle of data or rather the students don't necessarily have to understand the lifecycle, in order to manage the data. Another faculty cited this as why skills are lacking – students don't see the reasons why data management and organization are important. Another maintains that it is important for students to understand the entire process so that they can backtrack if a mistake is made.

The “learning-by-doing” approach is cited by many faculty as the environment where students obtain these skills. If students are not proficient in this area, several problems can arise, including code overwrites, haphazard organization, and the inability to locate specific data. Internships are cited as an example of where students obtain proficiency.

Data management plans are also described as very important, however that students should be able to follow them, rather than develop and create them.

## Student responses:

Almost all students admit that they learn data management skills through trial-and-error and standards for managing and organizing the data, if they exist, were learned through word-of-mouth. Several students mention that there are no formal policies in place for managing the data in the lab, and even those with lab policies are unaware of formal standards in the discipline. Organization of data is an issue recognized by the student for day-to-day tasks. For example it is sometimes difficult for one student to locate particular files and sometimes must go back and rerun coding to find the authoritative version.

In general, the students describe the processes of data management and not necessarily the reasons behind it. For example, most students keep copies of their data in multiple locations, but the ad hoc methods of saving create confusion rather than security.



Jake Carlson – PI (Purdue), Camille Andrews (Cornell), Marianne Bracke (Purdue), Michael Fosmire (Purdue), Jon Jeffries (Univ. of MN), Lisa Johnston (Univ. of MN), Megan Sapp Nelson (Purdue), Dean Walton, (Univ. of OR), Brian Westra (Univ. of OR), Sara Wright (Cornell)